NETFLIX MOVIES AND TV SHOWS CLUSTERING

Zeeshan Ahmad
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Netflix is one of the leading OTT platforms, not only in India but also internationally Netflix manages a large collection of TV shows and movies, streaming it anytime via online. The success of the OTT platforms depends on two things- the variety of content and appropriate recommendations to the users. This business is profitable because users make a monthly payment to access the platform. Exploratory Data Analysis is done on the dataset to get the insights from the information however the principal invalid qualities are taken care of. There are 12 features and around 7700 observations in the dataset and are mostly textual features. Clustering is a useful technique to achieve the best possible recommendations and increase the viewership of the platform

1. Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

2. Introduction

Netflix, Inc. is an American technology and media services provider and production company headquartered in Los Gatos, California. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. The company's primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.
Netflix is a popular entertainment service used by people around the world. This EDA will explore the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn. Netflix is known for its strong recommendation engines. They use a mix of content-based and collaborative filtering models to recommend tv shows and movies. In

this task, one can create a recommendation engine based on text/description similarity techniques.
I will proceed with reading the data, and then perform data analysis. The practice of examining data using analytical or statistical methods in order to identify meaningful information is known as data analysis. After data analysis, we will find out the data distribution and data types. We will train 3 Clustering ML algorithms to predict the output. We will also compare the outputs.

Let us get started with the project implementation.

3. Data descriptions

The dataset contains following columns:

Show id: Unique ID for every Movie / TV Show

type – Identifier - A Movie or TV Show

title – Title of the Movie / TV Show

director-director of the content

cast –Actors involved in the movie / show

country – Country where the movie / show was produced

date_added – Date it was added on Netflix

release_year – Actual Release year of the movie / show

rating – TV Rating of the movie / show

duration – Total Duration - in minutes or number of seasons

listed_in – genre

description – The Summary description

4. Steps involved

Handling missing values:

We will need to replace blank countries with the mode (most common) country.

It would be better to keep a director because it can be fascinating to look at a specific filmmaker's movie. As a result, we substitute the null values with the word 'unknown' for further analysis.

There are very few null entries in the date_added fields thus we delete them.

Duplicate Values Treatment:

Duplicate values do not contribute anything to accuracy of results. Our dataset does not contain any duplicate values.

Exploratory data analysis:

After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.

To get the understanding of the data and how the content is distributed in the dataset, its type and details such as which countries are watching more and which type of content is in demand etc. has been analyzed in this step.

The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.

Data Pre-processing:

Removing Punctuation:

Punctuations do not carry any meaning in clustering, so removing punctuations helps

to get rid of unhelpful parts of the data, or noise.

Removing stop-words:

Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Lemmatization:

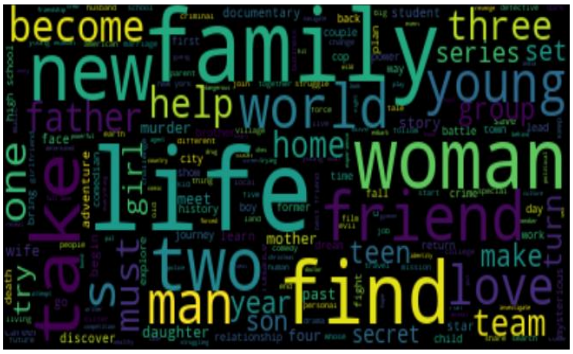
Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meanings to one word.

Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

Word cloud:

Word clouds (also known as text clouds or tag clouds) work in a simple way: the more a specific word appears in a source of textual data (such as a speech, blog post, or database), the bigger and bolder it appears in the word cloud.

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is.



Tfidf vectorization:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction. We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data. So, it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

5. Model Building

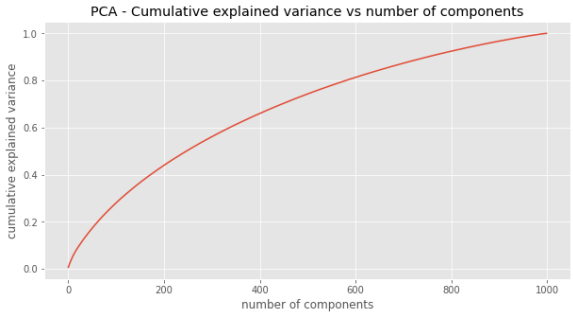
Dimensionality Reduction:

Principal Component Analysis is an unsupervised learning algorithm that is used

for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

It is a technique to draw strong patterns from the given dataset by reducing the variances.

PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.



Here we can spot that 80% variance is explained by 600 components only.

Clustering Algorithm:

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications

We have used 5 clustering algorithms:

1.K-Means Clustering

2. DBSCAN Clustering

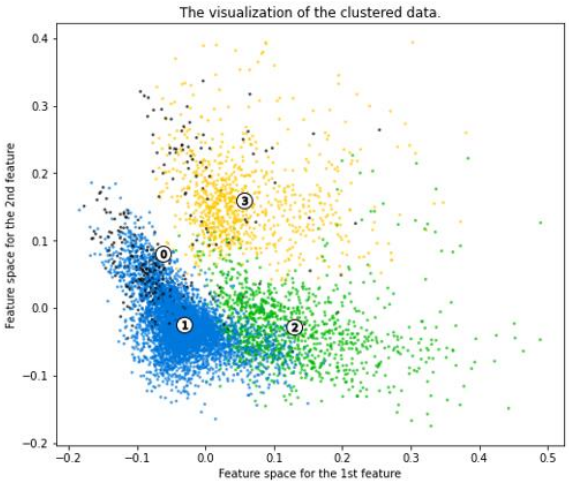
3. Hierarchical Clustering

K-Means Clustering:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

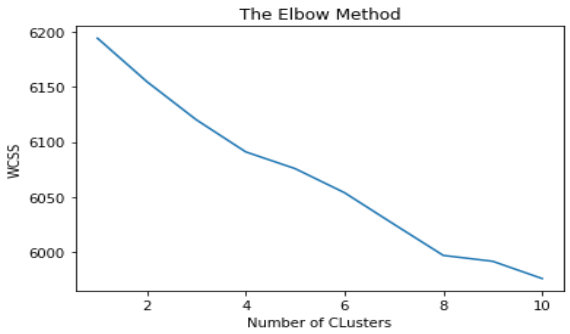


K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non overlapping subgroups where each data point belongs to only one group.

Elbow method:

The Elbow Curve is one of the most popular methods to determine this optimal value of k.



The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

Silhouette score:

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

Coefficient s for a single sample is then given as:

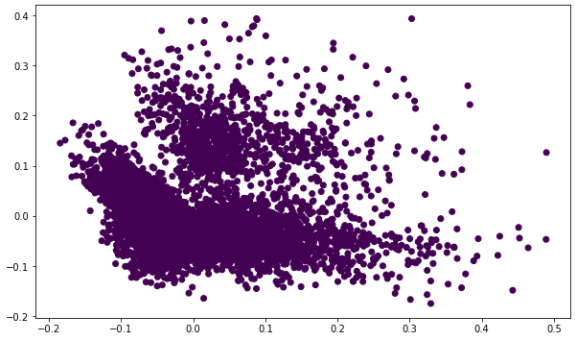
The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

```
cluster: 2      Sillhoute: 0.0065
cluster: 3      Sillhoute: 0.0091
cluster: 4      Sillhoute: 0.0079
cluster: 5      Sillhoute: 0.0101
cluster: 6      Sillhoute: 0.0109
cluster: 7      Sillhoute: 0.0116
cluster: 8      Sillhoute: 0.0124
cluster: 9      Sillhoute: 0.0128
cluster: 10     Sillhoute: 0.0124
```

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. ... a= average intra-cluster distance i.e., the average distance between each point within a cluster.

DBSCAN Clustering:

DBSCAN is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density.

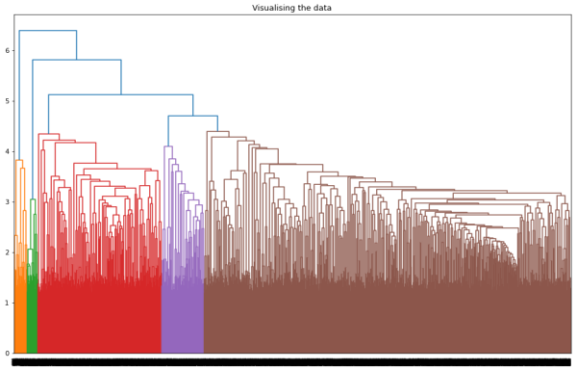


Given that DBSCAN is a density based clustering algorithm, it does a great job of seeking areas in the data that have a high density of observations, versus areas of the data that are not very dense with observations. DBSCAN can sort data into

clusters of varying shapes as well, another strong advantage.

Hierarchical Clustering:

Hierarchical clustering is separating data into groups based on some measure of similarity, finding a way to measure how they're alike and different, and further narrowing down the data.



A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

The number of clusters will be number of vertical lines which are intersected by the line drawn using the threshold

No. of Clusters = 4.

Conclusion

- In dataset there is around 69% content as movies and remaining 31% as TV shows.
- Most of contents are in TV-MA, TV-14 and TV-PG ratings for both Movies and TV Shows.
- Contents for TV Shows are more than movies in rating TV-Y and TV-Y7
- The content for children and general audiences is less in Netflix e.g., TV-Y7, TV-Y7-FV, G, etc.
- In year 2017 Count of Movies released are higher followed by 2018 and 2016.
- TV Shows are released higher in 2020 followed by 2019 and 2018.
- In year 2020 and 2021 TV shows are released more than movies
- There is a significant drop in the number of movies and TV Shows produced after 2020.
- The number of TV shows and movies added in 2019 and 2020 are maximum.
- International movies, dramas, and comedies are the top three genres with the most content on Netflix.
- United States have maximum movies in Netflix followed by India and United Kingdom.
- TV Shows are maximum in United States followed by United Kingdom and Japan.
- Content for adults are higher in United States followed by India, United Kingdom.
- Content for teens and kids are maximum in United States and India.
- Principal component analysis was performed in order to reduce the higher dimensionality
- Applied different clustering models K-means, hierarchical, DB Scan clustering on data we didn't get the best cluster arrangements.
- By applying different clustering algorithms to our dataset, we get the optimal number of clusters is equal to 4